

Research Article

Open Access

A Modified Hidden Markov Model and Its Application in Protein Secondary Structure Prediction

Sima Naghizadeh², Vahid Rezaeitabar², Hamid Pezeshk^{1*} and David Matthews³

¹School of Mathematics, Statistics and Computer Science, University of Tehran and Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

²Department of Statistics, Tarbiat Modares University, Tehran, Iran

³Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

Abstract

One of the important tools in analyzing and modeling biological data is the Hidden Markov Model (HMM), which is used for gene prediction, protein secondary structure and other essential tasks. An HMM is a stochastic process in which a hidden Markov chain called; the chain of states, emits a sequence of observations. Using this sequence, various questions about the underlying emission generation scheme can be addressed. Applying an HMM to any particular situation is an attempt to infer which state in the chain emits an observation. This is usually called posterior decoding. In general, the emissions are assumed to be conditionally independent from each other. In this work we consider some dependencies among the states and emissions. The aim of our research is to study a certain relationship among emissions, with a focus on the Markov property. We assume that the probability of observing an emission depends not only on the current state but also on the previous state and one of the previous emissions. We also use additional environmental information, and classify amino acids into three groups, using the Relative Solvent Accessibility (RSA). We also investigate how this modification might change the current algorithms for ordinary HMMs, and introduce modified Viterbi and Forward-Backward algorithms for the new model. We apply our proposed model to an actual dataset concerning prediction of the protein secondary structure and demonstrate improved accuracy compared to the ordinary HMM. In particular, the overall accuracy of our modified HMM, which uses the RSA information, is 63.95%. This is 5.9% higher than the prediction accuracy realized by using an ordinary HMM on the same dataset, and 4% higher than the corresponding prediction accuracy of a modified HMM that simply accounts for the dependencies among the emissions.

Keywords: Hidden Markov Models; Protein Secondary Structure Prediction; Relative Solvent Accessibility; Modified Viterbi Algorithm; Forward Algorithm

Abbreviations: ASA: Accessible Surface Area; C: Coil; DSSP: Dictionary of Secondary Structure of Protein; H: Helix; HMM: Hidden Markov Model; HSMM: Hidden Semi Markov Model; MHMM: Modified Hidden Markov Model; MSA: Multiple Sequence Alignment; PDB: Protein Data Bank; RSA: Relative Solvent Accessibility; RMHMM: Actual RSA; PRMHMM: Predicted RSA; S: Strand

Introduction

A hidden Markov model (HMM) is a statistical tool that is used to model a stochastic sequence. It corresponds to a Markov chain such that every state in the chain emits observations according to a density function. Using an HMM an observed sequence is modeled as the output of a discrete stochastic process, which is hidden. For each observation in the sequence, the process emits a symbol from a finite set of alphabets according to a probability density. HMMs are widely used in biological sequence analysis and bioinformatics. In particular, they are used in protein structure prediction studies. Hidden Markov Models first used in speech recognition problems, and typically involved mixture of Gaussian autoregressive densities which led naturally to a maximum likelihood solution of the familiar linear prediction analysis [1-4]. The first use of an HMM to predict protein secondary structure was published in 1993 [5]. A probabilistic model of protein sequence-structure relationship in terms of structural segments was proposed in [6], which formulated the question of secondary structure prediction as a general Bayesian inference problem. A Bayesian approach dealing with prior information on how to identify homogeneous segments was reported in [7]. A special type of HMM for labeled data was proposed

by [8] which developed a maximum likelihood method for estimating the parameters of the model. Won et al. [9,10] applied a new method for optimizing the topology of an HMM for the secondary structure prediction using genetic algorithms. They also applied an evolutionary method to optimize the structure of an HMM. HMMs and prior biological knowledge were combined in [11,12]. The use of an HMM with a reduced set of states for predicting protein structure was considered in [13]. The EM and the Viterbi algorithms for an HMM were implemented in linear memory in [14] and a position-specific HMM to predict protein structure was described by Cheng-Li et al. [15] that combined fragment assembly, clustering, target selection, refinement and consensus in one process. To reduce the scale of the HMMs for protein secondary structure prediction, Lee et al. [16] suggested a 9-state HMM. Karplus et al. [17] introduced a new HMM based server for protein structure prediction. He provided a large number of intermediate results, which are often interesting in their own right: multiple sequence alignments (MSAs) of putative homologs, prediction of local structure features, lists of potential templates of known structure, alignments to templates and residue-residue contact

***Corresponding author:** Hamid Pezeshk, School of Mathematics, Statistics and Computer Science, University of Tehran and Institute for Research in Fundamental Sciences (IPM), Tehran, Iran, E-mail: hamid.pezeshk@gmail.com

Received December 14, 2011; **Accepted** January 07, 2012; **Published** January 20, 2012

Citation: Naghizadeh S, Rezaeitabar V, Pezeshk H, Matthews D (2012) A Modified Hidden Markov Model and Its Application in Protein Secondary Structure Prediction. J Proteomics Bioinform 5: 024-030. doi:10.4172/jpb.1000209

Copyright: © 2012 Naghizadeh S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

predictions. Applying HMMs usually requires one to implement of the Viterbi algorithm [18,19].

Proteins are one of the most important molecules in any living cell and the study of protein structure is very important in biology. Proteins are polymer chains composed of 20 amino acids. Adjacent amino acids are connected with peptide bonds. For each amino acid, a secondary structure denotes the local spatial arrangement and the regularities of amino acids with respect to each other. Amino acids usually involve one of three main secondary structures: Helix (H), Strand(S) and Coil(C). Protein secondary structure prediction is an important problem in the field of bioinformatics, and predicting the secondary structure of a protein provides a starting point for predicting three-dimensional structure, which in turn enables researchers to identify a protein structure in its entirety. Various inferential techniques, such as the nearest neighbor method, neural networks, machine learning, and methods based on information theory have been used to predict protein structure. HMMs are also well known models for the problem of predicting the secondary structure of proteins.

In a few papers on the problem of predicting protein secondary structure, high-order dependencies among emissions or the corresponding states have been considered. For example, a new probabilistic method for protein secondary structure prediction based on dynamic Bayesian networks was reported in [20]. These investigators used a multivariate Gaussian distribution and were able to account for the dependency between the profiles and secondary structures as well as the dependency between profiles of neighboring residues. In [21] a high-order HMM that permits states and observations to depend on previous states were described. Aydin et al. [22] proposed a hidden semi-Markov model that accounts for the patterns of statistically significant amino acid correlation at structural segment borders. The same paper also introduced an alternative decoding technique for the hidden semi-Markov model (HSMM). The proposed method is based on the N-best paradigm, where a set of suboptimal segmentations (the N-best list) is computed as an alternative to the most likely segmentation [23].

Some papers concerning protein structure consider the environmental features, and thereby increase the accuracy of the predicted protein structure. The Relative Solvent Accessibility, RSA, is an aspect of protein analysis that has been widely studied. Various researchers have combined RSA information with HMMs to improve the precision of the protein structure prediction [24-29].

In what follows we discuss a modified HMM (MHMM) that allows for some dependencies among emissions. In an ordinary HMM, given the state, the emissions are assumed to be independent. In the modified model that we consider, we assume each emission depends not only on the current state, but also on the previous state and on the previous emissions. We also use the RSA of an amino acid to classify each residue into one of three groups. We call this model RMHMM. Compared to the ordinary HMM, our RMHMM improves the precision of predictions. We also investigate how our proposal RMHMM changes current algorithms for ordinary HMMs.

Material and Methods

Hidden markov models

An HMM is a stochastic process in which a hidden Markov chain of states emits a sequence of observations. A Markov chain is a sequence of random variables which has the Markov property; that

is, given the present state of the process, the future and the past states are independent. In ordinary Markov chain models, the sequences of states are observable, but in hidden Markov models these same states are not observable. Suppose we have an HMM model that involves N states. These states cannot be observed, but they emit some characters (alphabets) which are observable. In what follows, we consider some dependencies among the observable emissions. It is assumed that any emission depends on one of the previous emissions. In addition, we suppose that the next emission depends on the previous state of the hidden Markov chain. We call this model a modified hidden Markov model (MHMM). Let us introduce the notation for an MHMM through the following set of four specifications:

Let $S_t, t=1 \dots N$ denote a stationary Markov chain with N states and a transition probability matrix $A_{N \times N}$. Let $Q = q_1, q_2, \dots, q_T$ be a sequence of T consecutive states.

Then a_{ij} , element (i, j) of A is defined by

$$a_{ij} = P(Q_t = S_j | Q_{t-1} = S_i)$$

Where

$$\sum_{j=1}^N a_{ij} = 1.$$

We represent observed emission of the MHMM by $Y_t, t=1, \dots, M$. Suppose that the sequence Q of the states results in the observed emission $O = O_1, O_2, \dots, O_T$. We assume that given the emission Y_k at time $t-n$, the state S_i at time $t-1$ and the state S_j at time t , the probability matrix of emission Y_t at time t is denoted by

$$P = (p_{kl}^{ij}(n)),$$

Where

$$p_{kl}^{ij}(n) = \Pr(O_t = Y_i | O_{t-n} = Y_k, Q_{t-1} = S_i, Q_t = S_j).$$

Accordingly, the conditional probability distribution for an observable emission depends on the n^{th} previous emission as well as current and previous states.

Obviously

$$\sum_{i=1}^M p_{kl}^{ij}(n) = 1, i, j = 1, \dots, k = 1, \dots, M$$

4. If we denote the initial probability distribution of state S_i at time $t=1$ by π_i , we have

$$\Pi = \pi_1, \pi_2, \dots, \pi_T$$

Where

$$\pi_i = P(Q_1 = S_i)$$

The vector of probabilities corresponding to observable emission Y_k at time $t=1$ given state S_j at time $t=1$ is denoted by

$$B = (b_1(k), b_2(k), \dots, b_3(k)), k = 1, 2, \dots, M$$

Where

$$b_j(k) = P(O_1 = Y_k | Q_1 = S_j)$$

Let λ represent the entire parameter space for this MHMM.

Two of the most frequently asked questions corresponding to any Hidden Markov Model are:

Given an observed sequence of emissions from an HMM, how should we choose a corresponding sequence of states which is most probable?

Given an observed sequence of emission and the model, represented by λ , how can we calculate the probability of being in state S_i at time t ?

In what follows, we address these questions using our MHMM. Our goal is to show how the modified model can be constructed efficiently. In the process of applying the model to actual data, we also incorporate the information represented by RSAs of amino acids to see how these changes improve the performance of the MHMM with respect to the secondary structure prediction problem.

Finding the Most Probable State Sequence via the Viterbi Algorithm

For a given sequence of observed emission, we seek to find the most probable corresponding sequence of states. Suppose we observe the sequence

$$O = O_1, O_2, \dots, O_T$$

Assume that the corresponding sequence of states is

$$Q = q_1, q_2, \dots, q_T$$

So we want to find $\max_{Q \in \mathcal{Q}(\lambda)} p(O, Q | \lambda)$. The solution of this equation is identical to the solution of $\arg \max_{Q \in \mathcal{Q}(\lambda)} p(O, Q | \lambda)$. Given the model λ , the joint probability of observed emission and state sequences is

$$p(O, Q | \lambda) = p(O | Q, \lambda) p(Q | \lambda).$$

$$p(O, Q | \lambda) = \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} p_{q_1 q_2}^{q_1 q_2}(1) \dots$$

$$a_{q_{t-1} q_t} p_{q_{t-1} q_t}^{q_{t-1} q_t}(n) \dots a_{q_{t-1} q_t} p_{q_{t-1} q_t}^{q_{t-1} q_t}(n) \dots$$

Given a sequence of observed emissions, the Viterbi algorithm computes most probable path of hidden states, see for example, [4]. In what follows we introduce a modification of the Viterbi algorithm that can accommodate the dependencies in our MHMM.

A Modified Viterbi Algorithm

Using the notation of Rabiner et al. [4], we describe the modified Viterbi algorithm via the following specification

1. Initialization:

$$\delta_1(i) = \pi_i b_i(o_1),$$

$$\psi_1(i) = 0, \quad M_1(i) = 1, \quad i = 1, \dots, N.$$

2. Recursion:

$$M_t(j) = \max_i [\delta_{t-1}(i) a_{ij}], \quad t = 2, \dots, T,$$

$$\psi_j(t) = \arg \max_i [\delta_{t-1}(i) a_{ij}], \quad j = 1, \dots, N, \quad k = \psi_j(j).$$

It follows that

$$\delta_t(j) = \begin{cases} M_t(j) P_{O_{t-1}, O_t}^{j, j}(1), & t = 2, \dots, n, \\ M_t(j) P_{O_{t-1}, O_t}^{j, j}(n), & t = n+1, \dots, T, \end{cases} \quad j = 1, \dots, N.$$

Notice that in the first part of the expression (2), the first - order dependency among observed emissions is used. This is due to the lack of sufficient observed emissions at the beginning of the sequence to use the appropriate higher - order dependency.

3. Termination:

$$P^* = \max_i [\delta_T(i)], \quad q_T^* = \arg \max_i [\delta_T(i)],$$

4. State sequence backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \dots, 1.$$

Determining the state at time t via the Forward and Backward algorithm (the posterior decoding). For a sequence of observations, $O = O_1, O_2, \dots, O_T$ and a corresponding sequence of states, $Q = q_1, q_2, \dots, q_T$ our aim is to evaluate $p(q_t = S_j | O, \lambda)$. This step is usually called posterior decoding. In what follows, we show how the ordinary method can be modified to accommodate our MHMM.

Modification of the posterior decoding

Let us define the modified forward variable by:

$$\alpha'_{(i)}(ij) = P(O_1 O_2 \dots O_t, q_{t-1} = S_i, q_t = S_j | \lambda).$$

Where $\alpha'_{(i)}(ij)$ is the joint probability of $O_1, O_2, \dots, O_t$, q_{t-1} and q_t given the model λ . To calculate $\alpha'_{(i)}(ij)$ we proceed as follows:

1. Initialization:

$$\alpha'_{(2)}(ij) = \pi_i b_i(O_1) a_{ij} p_{O_1 O_2}^{ij}(1), \quad i = 1, \dots, N, \quad j = 1, \dots, N.$$

2. Induction:

$$\alpha'_{(t+1)}(jk) = \begin{cases} \sum_{i=1}^N [\alpha'_{(t)}(ij)] a_{jk} P_{O_t, O_{t+1}}^{jk}(1), & t = 2, \dots, n-1. \\ \sum_{i=1}^N [\alpha'_{(t)}(ij)] a_{jk} P_{O_t, O_{t+1}}^{jk}(n), & t = n, \dots, T. \end{cases} \quad j, k = 1, \dots, N.$$

The modified backward variable $\beta'_t(ij)$ can be defined in a similar fashion:

$$\beta'_t(ij) = P(O_{t+1} O_{t+2} \dots O_T | q_{t-1} = S_i, q_t = S_j, \lambda).$$

Hence we can proceed as follows to obtain $\beta'_t(ij)$:

1. Initialization:

$$\beta'_{(T)}(jk) = a_{jk} p_{O_T, O_T}^{jk}, \quad i = 1, \dots, N, \quad j = 1, \dots, N.$$

2. Induction:

$$\beta'_{(t)}(jk) = \begin{cases} \sum_{i=1}^N [\beta'_{(t+1)}(ik)] a_{jk} P_{O_t, O_{t+1}}^{jk}(1), & t = T-1, \dots, n+1. \\ \sum_{i=1}^N [\beta'_{(t+1)}(ik)] a_{jk} P_{O_t, O_{t+1}}^{jk}(n), & t = n, \dots, 2. \end{cases} \quad i, j = 1, \dots, N.$$

To obtain $p(q_t = S_j | O, \lambda)$, we first calculate

$$\gamma'_t(ij) = p(q_{t-1} = S_i, q_t = S_j | O, \lambda).$$

Obviously

$$\gamma'_t(ij) = \alpha'_{(t)}(ij) \beta'_t(ij) / p(O | \lambda).$$

If we sum over the possible values of S_i in expression (5), we obtain $p(q_t = S_j | O, \lambda)$

Next, we combine the MHMM with the information obtained from the RSA of each residue in a sequence of amino acids from the WHATIF dataset in the Protein Data Bank, PDB. We call the new model RSA combined with MHMM (or RMHMM).

Dataset and the order of dependency

To analyze the performance of our model, we applied our MHMM and RMHMM to a dataset of WHATIF PDB selection list that Momen-Roknabadi et al. [27] used in their work. The dataset contains 6970 chains with resolution ≤ 2.5 such that identity between each pair of sequences is not more than 30 percent. We have used this dataset for both training and testing, using 5 fold cross-validation. This is briefly discussed in results and discussion. Following Aydin et al. [23], we fitted our model to the above mentioned dataset, using various order of dependencies, i.e. $n=2,3,4$ for helices, $n=1,2$ for strands and $n=1$ for coils. So we may have 21 types of modeling. The results show that conditioning each emission on the second previous emission, yields the best performance with respect to prediction accuracy.

Protein secondary structure

Proteins are polypeptids (polymers) of amino acids. These polymers consist of 20 different amino acids. The secondary structure of a protein is the regular structure locally defined. These spatial regularities are due to hydrogen bonds between amino acids. The secondary structures are assigned to each amino acid using the DSSP program [30]. There are eight types of secondary structures defined by DSSP: G = 310 helix; H = α -helix; I = π -helix; T = hydrogen bonded turn; E = extended strand (β -sheet) conformation; B = residue in isolated β -bridge; S = bend; and amino acid residues which do not correspond to any of the above conformations are assigned as the eighth type 'Coil'. Although, the 8-state DSSP code already represents a simplification from the 20 amino acid residues that are present in a protein, the majority of secondary structure prediction methods further simplify matters to the three dominant states: helix (H), strand (S) and coil (C). The first two are periodic motifs that are characterized by geometrical features. The coil class is the default description for all amino acids that do not belong to the helix or strand classes. In the secondary structure prediction problem, we usually assign a structural state from a three-letter alphabet consisting of {H,S,C} to each amino acid.

Solvent Accessibility and RSA Prediction

The relative solvent accessibility degree determines whether a given amino acid is external or hidden. An amino acid is declared not exposed to solvent when its observed accessibility is less than a certain fraction of its observed accessibility in a reference state. Typical values for this threshold are around 20%. We use the ASA (Accessible Surface Area) from DSSP to determine the RSA of each residue by dividing the corresponding ASA value by the maximum possible ASA for each amino acid. Momen-Roknabadi et al. [27] used some different fixed RSA thresholds for all amino acids and the residue-specific RSA thresholds. They divided amino acids into two groups (buried and exposed) and three groups, (buried, intermediate and exposed) using the RSA values. They also used different thresholds like "mean RSA" and "median RSA" for binary classification and "mean $\pm$ standard deviation" and "first tertile-second tertile" for ternary classification. Following Momen-Roknabadi et al. [27] we apply a residue specific RSA threshold approach to classify amino acids into the three groups

known as buried, intermediate and exposed. According to their results, the best performance with respect to prediction accuracy is achieved when they use ternary classification, a residue dependent threshold, "mean $\pm$ standard deviation" cutoff and five-fold cross-validation. We use "mean standard deviation" of the RSA distributions as thresholds for the classification. Thus each amino acid has two cutoffs which are dependent on the mean and the standard error. Using this basis for classifying 20 types of amino acids results in 60 types of observations. We also used RVP-net [31] for predicting RSA values. The output of this program is an RSA value between 0% and 100%, which we then used to classify residues into the three classes; Buried, Intermediate, Exposed.

Results

An accuracy measure for evaluating the prediction

In the protein secondary structure prediction problem, one of the most commonly used measures of accuracy is

$$Q_3(\%) = 100(N_H + N_S + N_C) / N$$

where, N_H , N_S , and N_C represent the number of correctly predicted H, S and C state, respectively, and N is the total number of amino acids. In a similar fashion, we can also define

$$Q_k(\%) = 100N_k / N'_k, k = H, S, C,$$

Where N_k the total number of amino acids with is correctly predicted secondary structure of type k, and N'_k is the total number of amino acid of type k. The observed value of Q_k represents the sensitivity of the prediction.

Modified hidden markov model

In the typical HMM, the observed emissions are assumed to be independent of each other. However in this study we first adopt a modified HMM which permits some dependencies among the emissions. We assume that the probability of observing each emission depends not only on the current state but also on the previous state and on the previous emissions. This model can be used whenever dependency among emissions should be considered. To test if this assumed dependency is reasonable, we test our modified HMM on the protein secondary structure prediction problem. To implement our modified HMM on the dataset, we used the modified Viterbi and the modified Forward-Backward algorithms that we outlined in section 2. In order to compare the fit of our model with a classical HMM, we use various measures such as Q_3 and standard deviation, as well as five-fold-cross-validation [32]. Table 1 displays, Q_3 the percentage of secondary structures that are correctly predicted by each model. It also indicates the percentage of helix, Q_H strand, Q_S and coil, Q_C type, correctly predicted by HMM and MHMM using both the Viterbi and Forward-Backward algorithms. According to

Model	Viterbi				Forward-Backward			
	Q3	QH	QS	QC	Q3	QH	QS	QC
HMM	52.28	59.1	38.85	55.39	58.05	67.52	33.52	65.13
MHMM	54.24	60.53	43.65	56.35	59.87	71.24	43.45	59.98

Table 1: The accuracy of the protein secondary structure prediction for HMM and MHMM using Viterbi and Forward-Backward algorithms applying 5 fold cross-validation.

these tabulated results, the Q_3 criteria increases about two percent for both algorithms. The values of Q_H and Q_S also increase. Notice that, for the Forward-Backward algorithm, the value of Q_3 is roughly 10% greater for the MHMM than for the HMM. An improvement of this magnitude is important in the protein secondary structure prediction problem. These results clearly show that using modified Viterbi and Forward-Backward algorithms yields better prediction accuracies. Our results also suggest that some dependency between conformations and their influence on the secondary structures can be captured by our modified Viterbi and Forward-Backward algorithms. Table 2 shows the prediction accuracy of each amino acid when both algorithms are used. It is intriguing to note that predictions of almost all amino acids are improved by fitting an MHMM rather than an HMM. Figures 1 and 2, respectively, show bar charts of the improvement in prediction accuracy using MHMM and HMM for each amino acid, using the Viterbi and Forward-Backward algorithms.

The impact of solvent accessibility information on prediction

Environments around the protein residues can affect their propensity for different structures [28]. Therefore, amino acids may behave differently when they are located in the protein interior or on the surface. Based on this observation, other researchers have suggested that exploiting information concerning environmental factors, such as accessible surface area (ASA), might improve the prediction of

Viterbi		Forward-Backward		
Amino acid	HMM	MHMM	HMM	MHMM
A	56.03	57.69	61.68	61.56
C	53.15	55.29	58.7	59.75
D	47.1	49.6	53.07	54.33
E	48.77	51.53	53.93	57.18
F	52.47	55.19	57.79	59.45
G	51.29	54.78	57.14	58.15
H	51.31	54.89	56.94	59.41
I	61.46	57.18	67.97	68.1
K	47.94	50.46	54.06	57.03
L	47.81	50.27	52.9	55.82
M	50.27	51.99	55.36	57.56
N	47.8	49.95	53.91	56.71
P	59.69	61.04	65.41	66.61
Q	51.43	53.99	57.91	59.5
R	52.04	54.38	58.32	59.87
S	54.59	56.42	60.36	62.44
T	53.64	55.33	59.8	63.05
V	53.64	55.85	60.01	60.64
W	55.35	57.95	61.25	62.36
Y	49.78	51.04	54.53	57.85
TOTAL	52.28	54.24	58.05	59.87

Table 2: The accuracy of the protein secondary structure prediction for HMM, RMHMM and PRMHMM using Viterbi and Forward-Backward algorithms applying 5 fold cross-validation.

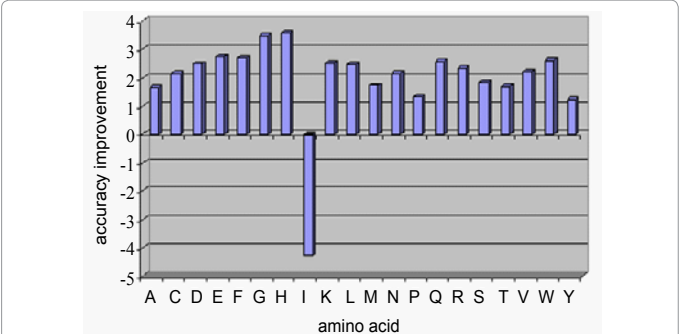


Figure 1: Improvements in Q_3 scores for MHMM compared to HMM using the Viterbi algorithm.

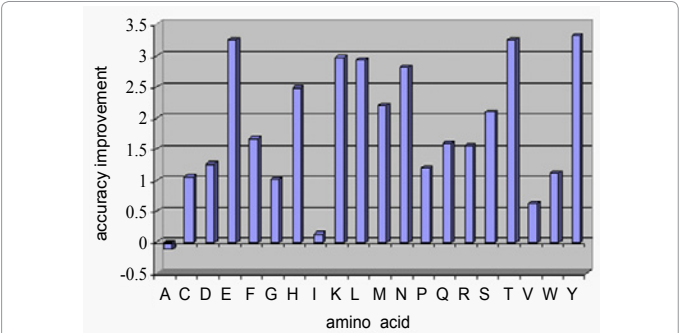


Figure 2: Improvement in Q_3 scores for MHMM compared to HMM using the Forward-Backward algorithm.

Viterbi Forward-Backward								
Model	Q3	QH	QS	QC	Q3	QH	QS	QC
HMM	52.28	59.1	38.85	55.39	58.05	67.52	33.52	65.13
RMHMM	57.07	58.66	54.6	59.1	63.95	67.33	57.3	67.39
PRMHMM	53.77	59.16	45.11	56	60.16	70.04	47.07	59.63

Table 3: The accuracy of the protein secondary structure prediction of each amino acid for HMM and MHMM using Viterbi and Forward-Backward algorithms applying 5 fold cross-validation.

secondary structures [25,27,29]. ASA is defined as the surface area of amino acid that is available to solvent. Relative solvent accessibility (RSA) is defined as the ratio of each amino acid that is accessible to solvent by dividing the corresponding ASA value to the maximum possible ASA. In this study we first derive the actual RSA from the DSSP program [30] and then use a residue - specific RSA threshold to classify each amino acid into one of three classes [27]. However, in practice we only know the sequence of the protein, and we should rely on the predicted RSA values rather than the actual values. Here we use the RVP-net predicted RSA [31]. Using the actual and predicted RSA, we develop two modified HMMs, called a Modified HMM using actual RSAs (RMHMM) and a Modified HMM using Predicted RSAs (PRMHMM), respectively. We use both the corresponding modified Viterbi and the Forward-Backward algorithms involving the actual and predicted RSA values to predict the secondary structure of proteins. Table 3 displays values of the criteria Q_3 and also Q_H , Q_S and Q_C using both algorithms for RMHMM and PRMHMM, and clearly demonstrates that using the actual and the predicted RSA values improve the prediction of secondary structures. Using RMHMM, the increases in Q_3 are about 4.8 % and 6% for the Viterbi and Forward-Backward algorithms, respectively. As expected, the impact of actual

RSA on prediction accuracy is greater than the effect of using predicted RSA.

Table 4 reports the observed prediction accuracy of each amino acid for HMM, RMHMM and PRMHMM using both the Viterbi and the Forward-Backward algorithms. In comparison with HMM, for almost all amino acids the secondary structures are predicted more accurately by both RMHMM and PRMHMM. When we use either modified model, using predicted RSAs, the observed accuracy of secondary structure predictions are lower than the corresponding values when actual RSAs are used. Figures 3 and 4 display the bar charts for prediction accuracy improvement for each amino acid in RMHMM, compared to HMM, using both the Viterbi and the Forward-Backward algorithms.

In general, as Tables 1-4 document, the prediction accuracy of RMHMM with respect to the goal of secondary structure prediction is superior to both HMM and MHMM. Moreover, our results suggest that corresponding both environmental information and dependency of emissions in HMM have considerable impact on the prediction accuracy of amino acid protein secondary structure. Compared to Momen-Roknabadi et al. [27], our results concerning the problem of secondary structure prediction show reasonable improvement. To test whether the difference between two values of Q_3 for HMM and RMHMM is significant, we propose to use the following statistical test.

Suppose X_i is the value of Q_3 for HMM, and Y_i is the corresponding estimate for in RMHMM, $i=1,\dots,k$ for which k is the number of iterations. We define

Viterbi Forward-Backward						
Amino acid	HMM	RMHMM	PRMHMM	HMM	RMHMM	PRMHMM
A	56.03	56.03	57.69	61.68	65.93	62.72
C	53.15	54.5	51.71	58.7	58.62	60.44
D	47.1	57.81	49.92	53.07	66	54.12
E	48.77	60.56	51.32	53.93	66.71	56.72
F	52.47	55.98	54.89	57.79	61.89	60.17
G	51.29	62.98	54.84	57.14	68.66	57.99
H	51.31	53.36	54.52	56.94	61.02	59.46
I	61.46	59.93	57.34	67.97	66.22	68.1
K	47.94	57.63	50.14	54.06	64.89	56.27
L	47.81	58.52	49.98	52.9	64.25	55.86
M	50.27	54.49	51.7	55.36	63.97	58.04
N	47.8	57.52	48.89	53.91	64.86	56.38
P	59.69	59.99	60.43	65.41	68.79	66.47
Q	51.43	57.82	53.78	57.91	64.69	60.28
R	52.04	56.48	53.57	58.32	64.27	61.08
S	54.59	53.81	56.01	60.36	61.18	62.89
T	53.64	53.93	55.26	59.8	60.7	63.29
V	53.64	59.39	55.89	60.01	65.23	62.13
W	55.35	53.38	57.85	61.25	59.97	63.4
Y	49.78	53.75	50.36	54.53	60.03	57.53
TOTAL	52.28	57.07	53.77	58.05	63.95	60.16

Table 4: The accuracy of the protein secondary structure prediction of each amino acid for HMM, RMHMM and PRMHMM using Viterbi and Forward-Backward algorithms applying 5 fold cross-validation.

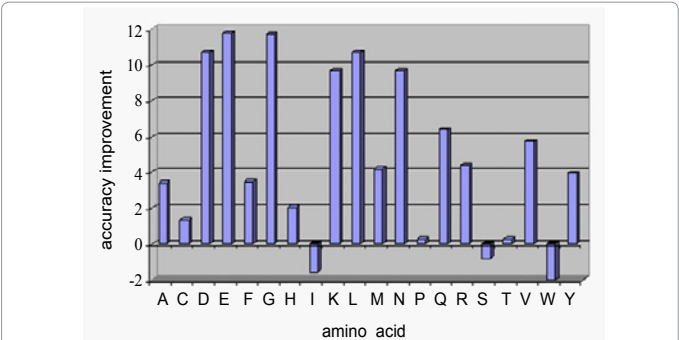


Figure 3: Improvements in Q_3 scores for RMHMM compared to HMM using the Viterbi algorithm.

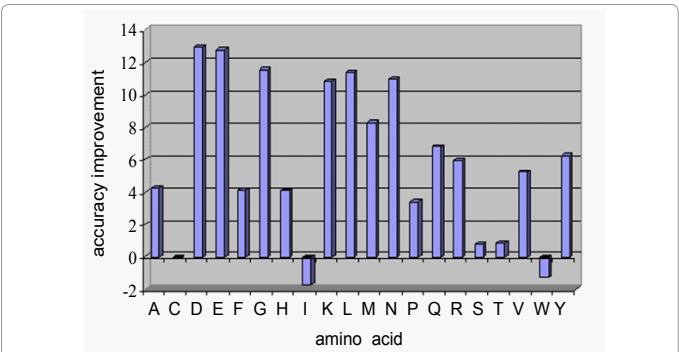


Figure 4: Improvements in Q_3 scores for RMHMM compared to HMM using the Forward-Backward algorithm.

$$Z_i = \begin{cases} 1 & Y_i > X_i \\ 0 & \text{otherwise} \end{cases}$$

So $P(Y_i > X_i) = P(Z_i = 1) = p$. The hypothesis of interest can be written as $H_0: p=1/2$, $H_1: p>1/2$. If we repeat the experiment n times independently, $T = \sum_{i=1}^n Z_i$ (i.e. the number of times that Y_i exceeds X_i)

is a random variable with a binomial distribution. Using the Neyman-Pearson test, for the significance level $\alpha=0.1$, $n=5$ (as already mentioned, we used five-fold cross-validation) and $p=1/2$, the test statistic becomes

$$\phi(t) = \begin{cases} 1 & t > 4 \\ 0.44 & t = 4 \\ 0 & t < 4 \end{cases}$$

If we suppose that Y_i is the value of Q_3 for RMHMM, we can use this test to determine whether the improvement for RMHMM is significant. The results show that in all five iterations, the Q_3 scores for both the MHMM and RMHMM are significantly greater than those for HMM. (See the additional files, which also contain information about Q_H , Q_S and Q_C).

Conclusions

Considering dependency among emissions seems reasonable and leads to some improvement in the prediction accuracy of protein secondary structure. Similarly, using RSA information will also result in improved prediction. So, combining RSA information with various dependency features in HMMs can be used in various hidden Markov models to improve the accuracy of predicting the protein secondary structure.

Acknowledgements

Hamid Pezeshk would like to thank the Department of Research Affairs of the University of Tehran. Sima Naghizadeh and Vahid Rezaeitabar are grateful to the Department of Statistics at Tarbiat Modares University. The authors would like to thank Mr. Amir Momen-Roknabadi for providing the dataset and Miss Nasim Ejlali for helpful comments on an earlier draft of this paper. Some parts of this work were carried out when Hamid Pezeshk was visiting department of statistics at Waterloo, Canada.

References

- Juang BH, Rabiner LR (1985) Mixture Autoregressive Hidden Markov Models for Speech Signals. *IEEE Trans Acoustics, Speech and Signal Processing* 33: 1404-1413.
- Juang BH, Rabiner LR (1991) Hidden Markov Models for Speech Recognition. *Technometrics* 33: 251-272.
- Rabiner LR, Juang BH (1986) An Introduction to hidden Markov models. *IEEE Trans Acoust, Speech Signal Process* 1: 257-286.
- Rabiner LR (1989) A tutorial on hidden Markov models and Selected application in speech recognition. In *Proceeding of the IEEE* 77: 257-286.
- Asai K, Hayamizu S, Handa K (1993) Prediction of Protein Secondary Structure by Hidden Markov Models. *Bioinformatics* 9: 141-146.
- Schmidler SC, Liu JS, Brutlag DL (2000) Bayesian Segmentation of Protein Secondary Structure. *J Comput Biol* 7: 233-248.
- Boys RJ, Henderson DA, Wilkinson DJ (2000) Detecting Homogeneous Segments in DNA Sequences by using Hidden Markov Models. *Applied Statistics* 49: 269-285.
- Krogh A (2001) Hidden Markov Models for Labeled Sequences. *IEEE Computer Society Press* 2: 140-144.
- Won KJ, Hamelryck T, Prugel-Bennett A, Krogh A (2007) An Evolutionary Method for Learning HMM Structure: Prediction of Protein Secondary Structure. *BMC Bioinformatics* 8: 357.
- Won KJ, Prugel-Bennett A, Krogh A (2004) Training HMM structure with Genetic algorithm for biological sequence analysis. *Bioinformatics* 20: 3613-3619.
- Martin J, Gibrat JF, Rodolphe F (2006) Analysis of an optimal hidden Markov models for secondary structure prediction. *BMC Struct Biol* 6: 25.
- Martin J, Letellier G, Marin A, Taly JF, de Brevern G, et al. (2005) Protein Secondary Structure Assignment Revisited: A Detailed Analysis of Different Assignment Methods. *BMC Struct Biol* 5: 17.
- Christos Lampros, Costas Papaloukasa, Themis P Exarchosa, Yorgos Goletsis, Dimitrios I Fotiadis (2007) Sequence-based protein structure prediction using a reduced state-space hidden Markov model. *Comput Biol Med* 37: 1211-1224.
- Churbanov A, Winters-Hilt S (2008) Implementing EM and Viterbi algorithms for Hidden Markov Models in linear memory. *BMC Bioinformatics* 9: 224.
- Cheng-Li S, Bu D, Xu J, Li M (2009) Fragment-HMM: a new approach to protein structure prediction. *Protein Sci* 17: 1925-1934.
- Sun Young Lee, Jong Yun Lee, Kwang Su Jung, Keun Ho Ryu (2009) A 9-state hidden Markov model using protein secondary structure information for protein fold recognition. *Comput Biol Med* 39: 527-534.
- Kevin Karplus (2009) SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res* 37: 492-497.
- Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13: 260-269.
- Forney GD (1973) The Viterbi Algorithm. *Processings of IEEE* 61: 268-278.
- Yao XQ, Zhu H, She ZS (2008) A dynamic Bayesian network approach to protein secondary structure prediction. *BMC Bioinformatics* 9: 49.
- Lee ML, Lee JC (2006) A Study on High-Order Hidden Markov Models and Applications to Speech Recognition. *LNCS* 4031: 682-690.
- Aydin Z, Yucel A, Borodovsky M (2006) Protein Secondary Structure Prediction For a Single-Sequence Using Hidden Markov Models. *BMC Bioinformatics* 7: 178.
- Aydin Z, Yucel A, Borodovsky M (2007) Bayesian Protein Secondary Structure Prediction With Near-Optimal Segmentations. *IEEE Transactions on Signal Processing* 55: 3512-3525.
- Zhu ZY, Blundell TL (1996) The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J Mol Biol* 260: 261-276.
- Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149: 445-458.
- Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 59: 467-475.
- Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi A (2008) Impact of residue accessible surface area on the prediction of protein secondary structure. *BMC Bioinformatics* 9: 357.
- Zhong L, Johnson WC (1992) Environment affects amino acid preference for secondary structure. *Proc Natl Acad Sci U S A* 89: 4462-4465.
- Macdonald JR, Johnson WC Jr (2001) Environmental features are important in determining protein secondary structure. *Protein Sci* 10: 1172-1177.
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577-2637.
- Ahmad S, Gromiha MM, Sarai A (2003) RVP-net: online predictions of real valued accessible surface area of proteins from single sequences. *Bioinformatics* 19: 1849-1851.
- Kohavi R (1995) A Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceeding of the Fourteenth International Joint Conference on Artificial Intelligence* 2: 1137-1143.